ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Speaker Recognition using MFCC and CORDIC Algorithm

Akhil Thomas[1]

P.G. Student, Department of Electronics Engineering, Viswajyothi Engineering College, Vazhakulam, Kerala, India[1]

ABSTRACT: Speaker recognition is a very important research area in speech synthesis and speech noise reduction. Speaker recognition is a new challenge for technologies. Many algorithms have been suggested and developed for feature extraction. This paper presents a feature extraction technique for speaker recognition using Mel Frequency Cepstral Coefficients (MFCC). Further, this paper evaluates experiments conducted along each step of the MFCC process. Finally the MFCC coefficients are retrieved using DCT. For simplicity, the DCT was done using CORDIC Algorithm. The results shows that the Speaker Recognition was accurate in this method.

KEYWORDS:MFCC, DCT, CORDIC Algorithm

I. INTRODUCTION

In recent years, there are so many mounting issues on Security. So, there is an increasing need for personal authentication using Information and Technology tools. In general, authentication for any person can be carried out in three different ways [1]: something the person has, something the person knows and something the person is. The first two methods of authentication have been carried out for several decades. However, they have shortcomings in that keys or credit cards can be stolen or lost. The third class of authentication methods is known as Biometric Authentication which has lesser problems compared to others. Each person has a unique anatomy and physiology which, if used for authentication purposes, provide safeguards against misuse.

Security systems can be improved by using voice. Speaker recognition refers to the task of recognizing people by their voices. The hidden commence for voice authentication is that every individual voice varies in pitch, tone, and volume enough to make it interestingly recognizable. Few elements contribute to this uniqueness; size and shape of the mouth, throat, teeth (articulators), and nose. The chance of having exactly the same unique combination of features between two people is slim.

Discrete Cosine Transform (DCT) is very complex on its calculations. So many techniques are available for doing DCT. Among them CORDIC algorithm is more simple because it contains lesser number of additions and multiplications. By this way the number of gates can be reduced and as a result of this the delay also decreases.

The main aim of this project is to realize a low cost Speech Recognition Module. The overall task is to design and implement a real time speaker recognition system using various digital signal processing tools in FPGA and MATLAB.

II. LITERATURE SURVEY

Speaker recognition is an important field in digital signal processing. A speaker recognition algorithm was proposed by [3] where MFCC was used to extract features and GMM to model and identify the accent of test speech signals. They reported that the MFCC-GMM based system shows 91 percent overall efficiency when compared to prosodic-NNC.

The feature extraction process for MFCC has been discussed in a number of works. The basic algorithm follows the pre-emphasis, framing, windowing, Fast Fourier Transform, Mel-filter bank processing and discrete cosine transform [4]. The final MFCC is the amplitude of the resulting spectrum. A study conducted on the improved Text-

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

independent Speaker Identification System for Real Time Applications [5] has divided pre-processing into three parts namely down sampling, Pre-emphasis and silence removal. Through pre-processing a reduced sampling rate, the desired changed amplitudes and noise removable signals are obtained. This can be used for obtaining a signal free of noise as a basis for efficient further processing.

A common approach for windowing is hamming window. Many papers have used hamming window for its windowing purposes. A paper on Biometric Identity Verification using Automatic Speaker Recognition [6] has explained the hamming window process. It shows that windowing each individual frame will minimize the spectral distortion to narrow the signal discontinuities for each frame both at the beginning and the end. Similarly, a study on Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector Quantization (VQ) Techniques [7] has also used hamming window technique. Their main focus was to observe a signal in finite time, when multiplied by a window function. Most of the papers only mentioned a rectangular window technique but do not offer experiments using this process. Rectangular window can be a good option as it smoothly truncates the ideal impulse response to avoid discontinuity [8].

III.SPEAKER IDENTIFICATION

Training (en-roll) phase and test (identification) phase are the two main process in Speaker Identification. For these both phases feature extraction is a common firststep, in which speaker dependent features are extracted fromthe speech sample. The feature extractor converts the digitalspeech signal into a sequence of numerical descriptors, calledfeature vectors. These feature vectors give a more steady,robust, and compact representation than the raw input signal.Thus this step motivates to diminish the amount of test datawhile holding the speaker discriminative information. In theen-roll phase Fig. 1, these features are stored in the speaker database.

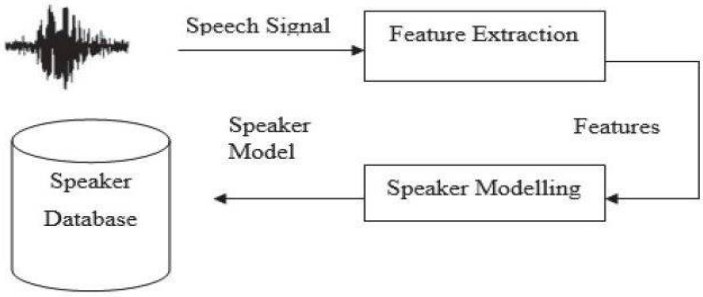

Fig. 1 En-roll Phase

The above figure shows the en-roll phase. It contains two blocks which is the feature extraction block and speaker modelling block. A speaker database is used for storing the speech signals. In this the speech signal was given to the feature extraction block and the features are to be modelled and the stores in the speaker database.

In the recognition phase Fig. 2, features are extracted from the unknown speaker's voice sample. The next phase is comparison with the speaker database. It is termed as pattern matching which refers to an algorithm that computes a match score between the unknown speakers feature vectors and the models stored in the database. The output of the pattern matching module is a similarity score.

Based on results obtained from the similarity score, the final decision about speaker identity is made. If the similarity score is equal or greater than a certain threshold value, the decision is the speaker number of the most similar speaker to the unknown speaker.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

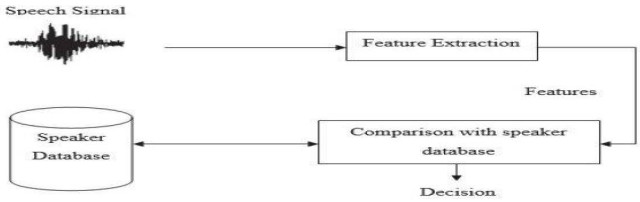

Fig. 2 Identification Phase

The Fig. 2 represents the Identification Process. This process also contains two blocks. They are the feature extraction and the comparison block. Here the features of the input signal is to be extracted in the first step. Then compare this values with old values in the data base. The comparison block will gives a score value.

If the similarity score between the unknown speakers feature vectors and all the speaker models in the database is less than the threshold value, then the test speaker is identified to be an unknown speaker. This is the last phase in the recognition chain.

IV. SPEAKER RECOGNITION USING MFCC

A block diagram of an MFCC process is given in Fig. 3. MFCCs are based on the known variation of the human ears critical bandwidths with frequency. They depend on the known proof that the information carried by low frequency components of the speech signal is phonetically more vital for humans than those conveyed by high frequency components. The method of computing MFCC is based on short-term analysis, and therefore from each frame a MFCC vector is computed.

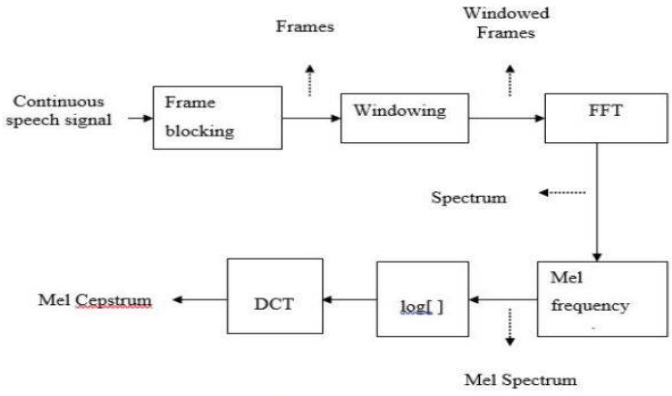

Fig. 3 MFCC Block Diagram

The Fig. 3 was the main block diagram of MFCC. It mainly contains 6 blocks. They are frame blocks, windowing block, FFT block, Mel frequency block, log block and DCT block. They are explained in detail on the following section

In this project, Mel Frequency Cepstral Coefficient is used for feature extraction. MFCC are coefficients that represent audio based on perception. They are derived from the Fast Fourier Transform of an audio clip to which Mel warping function has been applied. In this method, the frequency bands are positioned logarithmically, though in the Fourier Transform the frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in MFCC, it approximates the human system response more closely than any other system. The approximate formula for computing Mel for a given frequency f in Hz is

$$\text{Mel (f)} = 2595 * log_{10} (1+f/700)$$

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

a. Frame Blocking

Speech signals are fairly constant in short intervals of time. So they are processed in short time intervals. They are partitioned into frames with sizes generally between 20 to 30 milliseconds. Each frame overlaps its previous frame by a predefined size. The overlapping is done in order to smooth the transition from frame to frame. For frame blocking, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M. The frame blocking is done in order to reduce the complexity in processing the large signals.

b. Window Function

The next step is to window all the individual frames of the signal. This is done to minimize the unnatural signal discontinuities at the beginning and end of each frame. This is to select a portion of the signal that can reasonably be assumed to be stationary. A good window function has a narrow main lobe and low side lobe levels in their transfer functions. There is a trade-off between these two: making the main lobe narrower increases the side-lobe levels, and vice versa. The choice of the window is a trade-off between several factors.

c. FFT (Fast Fourier Transform)

DFT (Discrete Fourier Transform) can be computed via a faster algorithm called Fast Fourier Transform (FFT). A requirement for efficient FFT is that input signal has a length of 2M for some M power of two. In practice, the input signal is initially zero padded to the next highest power of two and the zero-padded signal is given as an input for the FFT. For example, if the length of the signal is 230 samples, it is zero padded to length N = 256 for which the FFT can be computed. Zeros can be added to the beginning or end of the signal, and it does not affect the result of the DFT. FFT reduces the computational complexity of DFT. Total complex multiplications and additions in FFT are (N/2.log2 N) and (Nlog2N) respectively compared toN^2complex multiplications and N(N-1) complex additions in case of DFT. The savings in computation time in practice are in the order of hundred folds. For instance, for N = 1024, the ratio of DFT multiplications to FFT multiplications is about 200 and the ratio of additions about 100.

d. Mel Frequency Conversion

The human ear perceives the frequency contents of the speech signal in non-linear fashion. Physiological research shows that the scaling is linear up to 1 kHz and logarithmic above that. The Mel-Scale (Melody Scale) filter bank characterizes how the human ear perceives frequency. It is used as a band pass filtering for this stage of identification. The signals for each frame are passed through Mel-Scale band pass filter to mimic the human ear. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. Mel is a unit of perceived fundamental frequency. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mel. Therefore, we can use the following approximate formula to compute the Mel for a given frequency f in Hz is,

$$\text{Mel}(f) = 2595 * log_{10}(1+f/700)$$

e. DCT using CORDIC Algorithm

This is the final step in feature extraction that leads to the generation of feature vectors. The conversion of log Mel spectrum into time domain using Discrete Cosine Transform (DCT) is carried out in this step. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called feature vectors. Therefore, each input utterance is transformed into a sequence of feature vector.The Coordinate Rotation Digital Computer algorithm (CORDIC) offers the opportunity to calculate all the desired functions in a rather simple and elegant way. The DCT based on CORDIC algorithm doesn't need multipliers. Moreover, it has regularity and simple hardware

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

architecture, which makes it easy to be implemented in VLSI. Also, the CORDIC based DCT algorithm can support the high performance applications such as HDTV due to high throughput. Also it will reduce the number of gates and thus the gate delay.

V. IMPLEMENTATION

The proposed work is implemented in MATLAB® 2013b using the toolbox VOICEBOX which is externally added into it. Signal Processing Toolbox is used for pre-processing, i.e., filtering and Fourier transform. The system is developed on Windows 10 64-bit operating system. The system is designed to recognize 7 different speakers.

The procedure for the entire process is given below,

1. Speaker 1 was spoke through the Microphone.
2. Speech was recorded as wav file.
3. Same process repeats for rest of the speakers.
4. The MFCC values for each speech signal was computed
5. During the test time only one speaker was spoke at a time
6. Then take the MFCC value of the test signal
7. Checks for the match between this value and already recorded MFCC values.
8. If a similarity was found, it will display the content. Otherwise the speaker was an unknown person.

VI. CONCLUSION

A basic speech recognition system which recognizes speech was thus developed. The system was trained in an environment with minimum ambient noise. One of the shortcomings of the system is that the performance degrades in the presence of noise. The system gives the most accurate results when implemented in the environment where it was trained. The system performance was analysed by increasing the number of training iterations. By using the MFCC feature extraction technique the accuracy of the detected speech was very high. Also the CORDIC Algorithm will reduces the number of gates and thus the gate delay must also decreases. From above all implementation and discussions it is clear that this project was a very good module for Speech Recognition.

REFERENCES

[1] D. Gibson, Understanding the Three Factors of Authentication Pearson IT certification, 2011.
[2] K. Kaur and N. Jain, "Performance analysis of text-dependent speaker recognition system based on template model based classifiers," IEEE Xplorer Digital Library, pp. 36-39, 2015.
[3] K. Mannepalli, P.N. Sastry, and M. Suman, "MFCC-GMM based accent recognition system for Telugu speech signals," Springer Link, pp. 87-3, 2015.
[4] Practicalcryptogrphy,(2013),Prcticalcryptography, [Online].http://practicalcryptography.com/miscellaneous/machine-learning/guide- mel-frequency-cepstral-coefficients-mfccs/
[5] M. Nagwa, M. Khalid AboElenein, M. A.Ibrahim, and M. Mohiy, "Improved text-independent speaker identification system for real time applications," IEEE Xplore Digital Library, pp. 58-62, 2016.
[6] K. Sujatha, R. P V Nageswara, R. A Arjuna, and Prasad K. Rajendra, "Biometric Identity Verification using Automatic Speaker Recognition," IEEE Xplorer Digital Library, pp. 1-5, 2015.
[7] J. Martinez, H. Perez-Meana, E. Escamilla, M. M. Suzuki, “Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques,” Research Gate, 2012. DOI: 10.1109/CONIELECOMP.2012.6189918.
[8] Andrew Greensted. (2010) The Lab Book Pages. [Online]. http://www.labbookpages.co.uk/audio/firWindowing.html.
[9] K.M Prabhu, Window Functions and Their Applications in Signal Processing, CRC Press, 2013.
[10] Georgina Brown“Automatic Accent Recognition Systems and the Effects of Data on Performance”Department of Language and Linguistic Science University of York, UK.
[11] Douglas Reynolds“Gaussian Mixture Models” MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA.
[12] M.A.Anusuya, S.K.Katti“Classification Techniques used in Speech Recognition Applications: A Review”Int. J. Comp. Tech. Appl., Vol 2 (4), 910-954.
[13] Ramesh Sridharan“Gaussian mixture models and the EM algorithm”.
[14] FadiBiadsy “Automatic Dialect and Accent Recognition and its Application to Speech Recognition”Submitted in partial fulllment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences.[20 Matthew Nicholas Stuttle, Hughes Hall“A Gaussian Mixture ModelSpectral Representation for Speech Recognition”Cambridge University Engineering Department